DOCUMENT RESUME

ED 290 771 TM 011 011

AUTHOR Kazelskis, Richard; And Others
TITLE The Original and Revised Nedelsky Procedure: Comparisons with Two Non-subjective Approaches to Determining Cutoff Scores.
PUB DATE Nov 87
NOTE 20p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (16th, Mobile, AL, November 11-13, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Competence; *Cutting Scores; Measurement Techniques; Methods; *Scoring Formulas; Standards; *Statistical Analysis; *Teacher Education; Test Construction
IDENTIFIERS *Nedelsky Method

ABSTRACT

Numerous techniques are available for determining cutoff scores for distinguishing between proficient and non-proficient examinees. One of the more commonly cited techniques for standard setting is the Nedelsky Method. In response to criticism of this method, Gross (1985) presented a revised Nedelsky technique. However, no research beyond that presented by Gross has yet to appear. This study examined and compared cutoff scores derived using the original and revised Nedelsky techniques and cutoff scores derived from two non-subjective standard setting techniques. Little evidence was found to suggest that the revised Nedelsky technique has substantially alleviated the problems associated with the original technique. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED290771

THE ORIGINAL AND REVISED NEDELSKY PROCEDURE: COMPARISONS WITH TWO NONSUBJECTIVE APPROACHES TO DETERMINING CUTOFF SCORES

Richard Kazelskis
James A. Siders
Mark G. Richmond
James O. Schnur
University of Southern Mississippi

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard Kazelskis
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 011 011

Paper presented at the annual meeting of the Mid-South Educational Research Association, Mobile, AL, 1987

Abstract

Numerous techniques are available for determing cutoff scores for distinguishing between proficient and nonproficient examinees. One of the more commonly cited techniques for standard setting is the Nedelsky method. In response to cricism of this method, Gross (1985) presented a revised Nedelsky technique. However, no research beyond that presented by Gross has yet to appear in the literature. The present study examined and compared cutoff scores derived using the original and revised Nedelsky techniques and cutoff scores derived from two nonsubjective standard setting techniques. Little evidence was found to suggest that the revised Nedelsky technique has substantially alleviated the problems associated with the original technique.

THE ORIGINAL AND REVISED NEDELSKY PROCEDURE: COMPARISIONS WITH TWO NONSUBJECTIVE APPROACHES TO DETERMINING CUTOFF SCORES

Determination of cutoff scores for distinguishing between proficient and nonproficient examinees is one of the more perplexing problems in measurement. Numerous procedures for determining performance standards are available in the literature (cf. Millman, 1973; Meskauskas, 1976; Hambleton, Swaminathan, Algina, & Coulson, 1978; Ebel & Frisbie, 1986; Lord, 1980; Hambleton & Swaminathan, 1985). Even compromise models are available (Beuk, 1984; De Gruijter, 1985). However, each approach results in a different cutoff score. Studies comparing the results of the different procedures and their characteristics are available (Andrew & Hecht, 1976; Glass, 1978; Harasyn, 1981; Halpin, Sigmon & Halpin, 1983; Norcini, Lipner, Langdon & Strecher, 1984). Unfortunately, there are no ultimate criteria for validating the standards defined by such procedures.

Several of the more commonly utilized standard-setting techniques require judgments of minimal competency by panels of subject matter experts to generate cutoff scores. These techniques include the Ebel method (cf. Ebel & Frisbie, 1986), the Angoff method (Angoff, 1971), and the Nedelsky method (Nedelsky, 1954). Also available are several techniques which are more objective in nature and are based upon various theoretical and statistical conceptions of test performance and/

or consideration of the consequences of errors in decision making. Several of these latter techniques are presented by Millman (1973), Meskauskas (1976), Hambleton et al. (1976), and Ebel and Frisbie (1986).

One of the first published standard-setting approaches was the Nedelsky method (Nedelsky, 1954). The Nedelsky technique has been criticised for a lack of a clear theoretical rationale, for the low correlations found between its item minimum pass indices (MPI) and the traditional measure of item difficulty, and because tended to give lower cutoff values than some of the other standard-setting techniques (Glass, 1978). In response to these criticisms Gross (1985) provided a revision of the technique which he felt alleviated some of these shortcomings. However, at present no further research into the characteristics of the revised Nedelsky technique appears available. Additional empirical support for the use of the technique is still needed.

The present study examines and compares cutoff scores derived using the original and revised Nedelsky techniques and cutoff scores derived from two nonsubjective standard setting techniques. The two nonsubjective techniques utilized were taken from Ebel and Frisbie (1986). The techniques were chosen because of their simplicity and because they are minimally dependent on normative data. The first technique only requires the number of test items and the number of choice options per item. This information is then used to determine the expected chance level score and the "ideal" mean test score with the cutoff score being

midway between the two values. The second nonsubjective technique used is similar in many ways to the first technique but incorporates the average and lowest test scores from a sample of test takers.

As Ebel (1979) points out, determining minimal performance standards involves making both arbitrary and not wholly satisfactory decisions. At present, continued empirical investigation into the statistical characteristics of the various procedures appears to be the only means for providing practitioners with information to help in selecting the technique which best fits their particular testing situation. The present study is an effort in this direction. It is aimed at providing additional empirical data to aid in the understanding of the characteristics of four standard setting techniques.

Methods and Results

Test Items

Two hundred ninety-five four option multiple choice items from the item pool of the Basic Core Examination (BCE) for the College of Education and Psychology at the University of Southern Mississippi were used. The items were constructed to measure student knowledge at the completion of the four courses constituting the basic teacher education core. The four courses include Educational Psychology, Public Education in the United States, Tests and Measurements, and the Psychology and Education of the Exceptional Child.

The items were constructed to measure performance relative

to 58 "indicators". The first 42 indicators are related to fourteen competencies defined by the Mississippi Teacher Assessment Instruments (MTAI). The remaining 16 indicators are related to course content tested by the National Teacher Examination (NTE). Scores can be generated for each of the four coursework areas as well as for three major components of the MTAI and concepts related to performance on the National Teachers Examination. Only scores related to the four coursework areas were utilized in the present study.

Subject Matter Experts

The group of subject matter experts used to obtain the Nedelsky values consisted of eight experienced public school teachers, one elementary school principal, and two teacher education faculty members. Their teaching experience ranged from three to twenty-six years with an average of 15.55 years. All of the subject matter experts were certified evaluators for the MTAI.

Standards-setting Techniques

Performance level cutoff scores were generated using the original Nedelsky technique, the Gross (1985) revised Nedelsky technique, and each of two nonsubjective techniques. To obtain the minimum pass index (MPI) for each item required for the Nedelsky techniques, the subject matter experts were asked to indicate which of the answer choices a minimally competent student should be able to eliminate as being incorrect. The possible MPI values for both the original and revised Nedelsky

techniques for four choice test items are presented in Table 1.

The two nonsubjective techniques utilized in the present study are presented in Ebel and Frisbie (1986). Procedure 1 (P_1) defines the minimum passing (MP) score for a test consisting of N items of k choices as $MP_1 = [N(k + 3)]/4k$. In words, MP_1 is obtained as follows: (1) determine the expected chance score (N/k), (2) obtain the ideal mean scores as midway between N and N/k, and (3) define MP_1 as the score value midway between the ideal mean and the expected chance score. When expressed as a percentage value MP_1 becomes independent of the number of items and depends only on k, i.e., for sets of items all having the same number of answer choices MP_1 becomes a constant. For k = 4 MP_1, expressed as a percentage, is 43.75.

The second nonsubjective technique (P_2) utilizes the mean test score (M) and the lowest obtained score (L) in defining MP. For procedure 2, $MP_2 = [2k(M + L) + N(k+3)]/8k$. In words, the procedure 2 MP is obtained by (1) determing the the midpoint between the expected chance score and the lowest obtained score, (2) determining the midpoint between the actual mean test score and the ideal mean (as defined in the preceding paragraph), and (3) defining MP_2 as the score value midway between the values obtained in steps 1, and 2.

Procedures

In order to reduce the number of items to be evaluated by the subject matter experts, the original item pool was randomly split, within indicators, to create two test forms. In that

different numbers of items were available for each indicator, the randomization resulted in differing numbers of items for the two forms. Form A consisted of 149 items, and Form B consisted of 146 items. Six of the subject matter experts were assigned to evaluate the form A items. The remaining 5 experts rated the form B items. MPI's were determined for each item using the original and revised Nedelsky techniques. Mean MPI's were then calculated for each item by averaging across subject matter experts. The cutoff scores, expressed as a proportion, were then determined by summing the item mean MPI's. For statistical analysis, all cutoff scores were expressed as percentage scores.

Results

The data were analyzed using a 2 X 3 (rating group/form by procedure) split-plot analysis of variance. A test for sphericity (Kirk, 1982) was run, and the null bypothesis was accepted ($p = .09$). Significant main effects were found for rating group/form ($p < .05$) and for standard setting procedure ($p < .001$), and a significant interaction ($p < .001$) was found between rating group/form and procedure (Table 2).

Analysis of the procedure main effect using the Newman-Keuls technique indicated significantly higher mean cutoffs for the revised Nedelsky procedure than for either the original Nedelsky or procedure two ($p < .01$). No significant difference was found between the original Nedelsky and the objective procedure two means.

Additional comparisons of the three procedure main effect

means with the constant for objective procedure one (P_1) indicated P_1 to result in a significantly lower cutoff score than either of the other three methods.

The significant main effect for rating group/form was the result of a higher mean cutoff score for group/form B. Due to the confounding of test form and rating group, and the presence of the significant group/form by procedure interaction, interpretation of this effect is deferred in favor of clarification of the interaction.

A graphical representation of the group/form by procedure interaction is presented in Figure 1. Comparisons among the cell means indicated no significant differences in the P_2 cutoff scores across the two test forms. However, significant differences ($p < .001$) were found between the mean cutoff scores generated by the two rating groups on both the original and revised Nedelsky procedures. Very different results were found among procedures within the two levels of the rating group/form dimension. Within the rating group/form A level, the original Nedelsky procedure resulted in a significantly lower ($p < .05$) cutoff score than did the revised Nedelsky procedures, and in turn, P_2 resulted in a signficantly higher ($p < .05$) mean cutoff score than did the revised Nedelsky procedure. However, within the rating group/form B level, no differences were found between the original and revised Nedelsky procedures which were each significantly lower than the mean cutoff scores for P_2.

Additional comparisons between the cell means and the P_1

cutoff value indicated that the mean cutoffs for the original and revised Nedelsky procedure were significantly higher than that found by P_1 with no significant difference in the cutoffs for the two objective techniques within group/form B. However, within group/form A the P_1 cutoff was signficantly higher than the mean cutoff for the original Nedelsky procedure, significantly lower than that found through P_2, and not significantly different from that found for the revised Nedelsky procedure.

Finally, correlations between the original and revised Nedelsky item values and item difficulties (proportion of test takers answering the item correctly) were examined. As found in previous studies, these correlations were disappointingly low and generally nonsignificant (Table 4). Only four of the sixteen correlations reached significance at the .05 level. The maximum correlation found was .43. The Gross revision of the original Nedelsky procedure did not improve the magnitude of the correlations. Two of the four significant correlations were found for the original Nedelsky procedure, and four were found for the revised Nedelsky procedure. In all instances little differences were found between the correlations found with the two procedures.

Discussion

The results suggest that Gross (1985) succeeded partially, at best, in his attempt to alleviate some of the weakness found in the Nedelsky technique. The revised Nedelsky procedure resulted in a higher mean cutoff score than did the original

procedure, when averaged across the rating groups. However, within rating groups, the revised procedure resulted in a significantly higher mean cutoff for one of the rating groups but not the other. Furthermore, the correlations between the item indices for the revised Nedelsky and item difficulties were not improved by the Gross revision.

Of primary concern is the variability in the cutoffs generated by the Nedelsky techniques across the two subject matter expert groups. This is in contrast to the quite similar cutoff scores found using P_3. Differences in cutoffs of 10 to 15 percentage points were found between the Nedelsky values for the two rating groups. In that the two test forms represent a random split of the original item pool, the differences in the Nedelsky values, both original and revised, are quite likely due to differences in the two rating groups perceptions of minimum competency. Determination of what constitutes minimum competency was left up to the individual groups. Prior to examination of the test items, each group was allowed to discuss the concept of minimum competency with the hope of reaching a consensus. Possibly the discussions led to differing perceptions of the concept.

Although explanation of the variability in the cutoff scores generated by the rating groups may be due to differences in perceptions of minimum competency, it could also be, as Halpin et al. (1983) suggest, due to a lack of understanding of the process of eliminating choice options. Halpin et al. found that the most

divergent cutoff scores for both the Nedelsky and the Angoff techniques were produced by school teachers, while university faculty and graduate students produced quite similar cutoff scores. Eight of the eleven raters used in the present study were public school teachers. Perhaps, school teachers do not constitute a viable source of subject matter experts for the Nedelsky technique when determining cutoff scores for college level tests.

With the variability in cutoff scores between the two rating groups, it is difficult to determine their merits relative to the two nonsubjective procedures. However, the differences in the cutoffs generated by the two procedures were not nearly as extreme as those within or between the two Nedelsky procedures. P_1, which is midway between the "ideal" mean and the chance score, appears to result in somewhat of a lower bound for the cutoffs generated by the other three approaches. P_2 tended to generate cutoff scores which were slightly higher than those produced by P_1, and in general quite similar to, though less variable than, those produced by the original Nedelsky procedure. As such, P_2 also tended to produce cutoffs which were somewhat lower than those produced by the revised Nedelsky technique.

In general little evidence was found to support the Nedelsky procedure either in its original or revised form. Particular concern is warranted if school teachers are to be utilized in determining cutoff scores with the Nedelsky techniques, at least for college level tests. Care should be taken to ensure that any

group of raters has a clear and concise concept of minimum competency for the referent group and a grasp of the concepts involved in the use of the Nedelsky techniques. Without such precautions the derived cutoff scores may be group specific.

References

Andrew, B. J. & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council on Education.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147-152.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. Journal of Educational Measurement, 22, 263-269.

Ebel, R. L. (1979). Essentials of educational measurement, (3rd ed.). Englewood Cliffs, N. J.: Prentice-Hall.

Ebel, R. L. & Frisbie, D. A. (1986). Essentials of educational measurement. (4th ed.). Englewood Cliffs, N. J.: Prentice-Hall.

Glass, G. V. (1978). Standards and criterion. Journal of Educational Measurement, 15, 237-261.

Gross, L. J. (1985). Setting cutoff scores on credentialing examinations: A refinement in the Nedelsky procedure. Evaluation and the Health Pofessions, 8, 469-493.

Halpin, G., Sigmon, G. & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hambleton, R. K., Swaminathn, H., Algina, J, & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48, 1-47.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Klewer-Nijhoff.

Harasyn, P. H. (1981). A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcomes. Educational and Psychological Measurement, 41, 225-234.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Monterey, Cal.: Brooks/Cole.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N. J.: Lawrence Erlbaum.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 46, 133-158.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 43, 205-216.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Norcini, J. J., Lipner, R. S., Langdon, L. O. & Strecher, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Table 1

Possible Minimum Pass Index (MPI) Values for the Original and Revised Nedelsky Procedures

No. of Viable Distractors	Original Nedelsky	Revised Nedelsky
1	1.0000	.8750
2	.5000	.5833
3	.3330	.4375
4	.2500	.3500

Table 2

Summary of Analysis of Variance

Source	df	MS	F	p	Greenhouse-Geisser p	Huynh-Feldt p
Group/Form (G)	1	950.04	41.33	<.001		
Sub/Grps	6	22.99				
Procedure (P)	2	96.50	15.58	<.001	<.005	<.002
G X P	3	468.67	75.66	<.001	<.001	<.001
P X Sub/Grps	12	6.19				

Table 3

Means and Standard Deviations of Cutoff Scores by Group/Form and Procedure

Group/ Form		Original Nedelsky	Revised Nedelsky	P_z	Row Mean
A	Mean	35.50	44.75	50.50	43.94
	S.D.	4.65	2.99	4.43	
B	Mean	59.75	63.00	45.75	53.25
	S.D.	1.50	1.41	4.03	
Col.	Mean	47.63	53.88	48.13	

Table 4

Correlations Between the Nedelsky MPI's and Item Difficulty

Area	Form	Original Nedelsky	Revised Nedelsky	df
PSY	A	.05	.04	29
	B	.43*	.43*	29
REF	A	-.05	-.01	35
	B	.27	.27	34
SPE	A	.31	.32	30
	B	.03	.01	30
T&M	A	.32*	.29*	47
	B	.07	.07	42

* $p < .05$

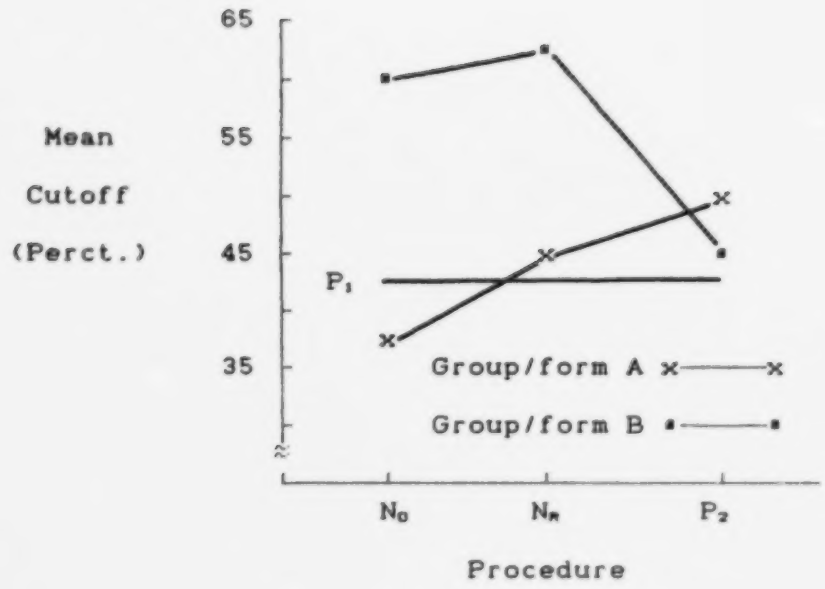

Figure 1. Mean Cutoff Score by rating group/form and procedure.